DOI: 10.5281/zenodo.7539922
DRAFT

Transforming Research through Innovative Practices for Linked Interdisciplinary Exploration

[31 December 2022] Advancing Open Scholarship

D2.4 – Report on identification and creation of new vocabularies

Version 1.0 –PUBLIC

H2020-INFRAEOSC-2019
Grant Agreement 863420

DRAFT

The project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 863420

Disclaimer- "The content of this publication is the sole responsibility of the TRIPLE consortium and can in no way be taken to reflect the views of the European Commission. The European Commission is not responsible for any use that may be made of the information it contains."

This deliverable is licensed under a Creative Commons Attribution 4.0 International License

D2.4 – Report on identification and creation of new vocabularies

Project Acronym:	**TRIPLE**
Project Name:	**Transforming Research through Innovative Practices for Linked Interdisciplinary Exploration**
Grant Agreement No:	**863420**
Start Date:	**1/10/2019**
End Date:	**31/03/2023**
Contributing WP	**WP2 Data acquisition**
WP Leader:	**IBL-PAN**
Deliverable identifier	**D2.4**
Contractual Delivery Date: 12/2022	**Actual Delivery Date: 01/2023**
Nature: Report	**Version: 1 Final**
Dissemination level	**PU**

DRAFT

Revision History

Version	Created/Modifier	Comments
0.0	Haris Georgiadis;EKT, Marta Blaszczynska;IBL-PAN, Maciej Maryl;IBL-PAN	Draft version
0.1	Leonie Disch; Know-Center, Julien Homo;FoxCub	Review
1.0	Haris Georgiadis;EKT	Final version

DRAFT

Table of Contents

1 Definitions **6**

2 Methodology **8**

2.1 Choosing LCSH as the base vocabulary 8

2.2 Selecting a reasonable number of SSH concepts from LCSH 13

2.3 Using Semantics.gr as the hosting platform 14

2.4 Making the GoTriple Vocabulary multilingual by leveraging existing mappings to LCSH 16

2.4.1 Adding labels from skos:closeMatch links of LCSH to other vocabularies 16

2.4.2 Adding labels from the National Library of Florence Vocabulary 17

2.4.3 Adding labels from the Rameau Thesaurus 17

2.4.4 Adding labels from the ELSST thesaurus 17

2.4.5 Validating preferable labels obtained by "skos:closeMatch" mappings 18

2.4.6 Language coverage after processing all available mappings 19

2.5 Increasing the multilingualism of the GoTriple Vocabulary using automatic translation 20

2.6 Adding new concepts from LCSH 22

3 The GoTriple Vocabulary - An Overview **25**

3.1 The SKOS data model 25

3.2 Overview of the Vocabulary 26

3.3 Export in RDF and Linked Data endpoint 31

4 APPENDIX - Maintenance and Sustainability **33**

5 References **37**

DRAFT

Table of Figures

Figure 1. The landing page of the 1st version of the GoTriple Vocabulary in semantics.gr 16

Figure 2. A google sheet for French where the a colleague from openedition validated prefLabels obtained by "skos:closeMatch" mappings 13

Figure 3. Adding more labels for a particular concept using the edit form of Semantics.gr 20

Figure 4. A Google spreadsheet used for validating french translations: validating a translation by ticking the

respective checkbox 21
Figure 5. The GoTriple Vocabulary landing page in semantics.gr 26
Figure 6. Top-level concepts 27
Figure 7. Unfolding concepts to navigate through its narrower concepts. 28
Figure 8. The concepts of the vocabulary in list view 28
Figure 9. Searching concepts by word or phrase. 29
Figure 10. A concept view page. 30
Figure 11. The edit form to modify or create a new concept. 31
Figure 12. The SKOS RDF Representation (in RDF/XML serialisation) of a concept. 32

Table of Tables

Table 1. Cartography of existing vocabularies 8
Table 2. mapping from SSH fields of science (FOS) of Frascati to LCSH 13
Table 3. The coverage per language after importing labels obtained by existing mappings 19
Table 4. Coverage per language after importing labels from automatic translation 21
Table 5. The coverage per language after importing 810 new concepts 24
Table 6. The SKOS properties used in the GoTriple Vocabulary 25

Acronyms

EKT	National Documentation Centre in Greece
LCSH	Library of Congress Subject Headings
LOD	Linked Open Data
RDF	Resource Description Framework
SKOS	Simple Knowledge Organization System
SSH	Social Sciences and Humanities

Publishable Summary

DRAFT

The GoTriple platform is a discovery service for SSH publications. It can be classed as an aggregator since it harvests publication metadata records from distributed sources (namely other aggregators or repositories). During the ingestion pipeline, it transforms metadata records into the Triple Data Model, it performs a series of cleansing, normalisation and enrichment procedures - in order to deal with metadata heterogeneity, increase multilingualism and improve content searchability and discoverability - and, finally, it stores and indexes the enriched metadata records, making them searchable via the GoTriple search engine.

Two of the most important enrichment procedures that metadata records undergo are classification and annotation. The former uses machine learning technology to automatically classify each publication using the MORESS classification scheme (D2.3[1]). The latter searches specific metadata fields of a record (titles, descriptions/abstracts and subjects/keywords) to assign them concepts from a multilingual LOD vocabulary of SSH concepts. The record is then updated with the respective links (concept URIs) to the concepts, as well as all available labels in the different languages. We call a concept URI with all the available labels that we add to a metadata record an annotation or Triple Keyword. Triple keywords are distinguished from the subjects/keywords of the original metadata. Since objects are indexed with annotation labels in all available languages, they are found when a search term matches an annotation label in any of the available languages. This way, both searchability and multilingualism are increased.

This deliverable describes the work and presents the outcome of task T2.4 “Cartography and creation of new vocabularies”. The objective of the task was to create a vocabulary of SSH concepts with labels in the 10 languages supported by the annotation service. The outcome is the GoTriple Vocabulary, a multilingual hierarchical set of 3,375 SSH-related concepts. It is a subset of LCSH (Library of Congress Subject Headings) that covers popular SSH subject areas. The English labels are enhanced with labels in Greek, French, Polish, German, Italian, Portuguese, Spanish, Croatian and Ukrainian. The vocabulary conforms to the SKOS data model and is published as Linked Open Data (LOD) under http://semantics.gr/authorities/vocabularies/SSH-LCSH in Semantics.gr, which is a platform developed by EKT for managing and publishing LOD vocabularies, thesauri and authority files of any schema. The vocabulary is used by the annotation service but, at the same time, is a standalone product, since it is published under an open license and can be used by the SSH research communities. The biggest challenges we faced in creating the vocabulary were a) choosing a base vocabulary b) defining a reasonable number of SSH concepts and c) adding labels in all GoTriple languages .

DRAFT

[1] https://drive.google.com/file/d/1K-83AHQdrfIhBEE2X84lRtnTTRq19gRp/view?usp=share_link

1 Definitions

Vocabulary

A collection of semantic resources (or terms) usually specific to a subject domain and created following a particular methodology. The semantic resources of a vocabulary are frequently organised hierarchically and may be interlinked forming a whole network of terms.

The term 'Vocabularies' comprises the following types:

Thesauri: vocabularies of terms that are usually domain-specific, organised hierarchically in one or more levels, and containing synonyms, antonyms, etc (ie Getty AAT[2])

Classifications and classification schemes: this type of vocabulary presents terms in thematic categories. Classification schemes often contain serial alphanumeric values (ie DDC[3], LCC[4], the MORESS classification[5])

Authority files: list of standardised terms for person names, corporate bodies or concepts commonly used within a domain, ie by libraries and memory institutions for cataloguing purposes. They are mainly based on MARC documentation models. Some examples: VIAF[6], LCSH, MeSH[7].

Semantic Resource or term

An informational resource that belongs to a vocabulary and which represents or describes a concept, a timespan, a place, a person or a corporate body. A semantic resource, or term includes a persistent URI, a preferable label - sometimes in more than one language - and a series of values that correspond to different properties. A semantic resource is an instance of one or more classes and assigns values to the individual properties.

Linked Open Data

Linked Data is a set of design principles for sharing machine-readable interlinked data on the Web based on the RDF standards of the semantic web. When combined with Open Data (data that can be freely used and distributed), it is called Linked Open Data (LOD)[8].

DRAFT

[2] https://www.getty.edu/research/tools/vocabularies/aat/
[3] https://en.wikipedia.org/wiki/Dewey_Decimal_Classification
[4] https://www.loc.gov/catdir/cpso/lcc.html
[5] https://project.gotriple.eu/wp-content/uploads/2020/12/MORESS-categories-for-GOTRIPLE-2020_12.pdf
[6] https://viaf.org/
[7] https://www.nlm.nih.gov/mesh/meshhome.html
[8] https://www.ontotext.com/knowledgehub/fundamentals/linked-data-linked-open-data/

Data Model (in the context of Linked Data and RDF)

A group of classes, properties and relationships that can be used to describe information resources in a particular knowledge domain. Examples: SKOS, Schema.org, Europeana Data Model, the Triple Data Model, DCMI Metadata Terms.

RDF

RDF is a standard model for data interchange on the Web [2]. RDF has features that facilitate data merging even if the underlying schemas differ, and it specifically supports the evolution of schemas over time without requiring all the data consumers to be changed.

RDF extends the linking structure of the Web to use URIs to name the relationship between things as well as the two ends of the link (this is usually referred to as a "triple"). Using this simple model allows structured and semi-structured data to be mixed, exposed, and shared across different applications.

This linking structure forms a directed, labeled graph, where the edges represent the named link between two resources, represented by the graph nodes. This graph view is the easiest possible mental model for RDF and is often used in easy-to-understand visual explanations[9].

SKOS

Simple Knowledge Organization System (SKOS) is a W3C recommendation designed for representation of thesauri, classification schemes, taxonomies, subject-heading systems, or any other type of structured controlled vocabulary [3]. SKOS is part of the Semantic Web family of standards built upon RDF and RDFS, and its main objective is to enable easy publication and use of such vocabularies as linked data.

The SKOS defines the classes and properties sufficient to represent the common features found in a standard thesaurus. It is based on a concept-centric view of the vocabulary, where primitive objects are not terms, but abstract notions represented by terms. Each SKOS concept is defined as an RDF resource. Each concept can have RDF properties attached, including:

- one or more preferred index terms (at most one in each natural language)
- alternative terms or synonyms
- definitions and notes, with specification of their language

Concepts can be organised in hierarchies using broader-narrower relationships, or linked by non-hierarchical (associative) relationships. Concepts can be gathered in concept schemes, to provide consistent and structured sets of concepts, representing whole or part of a controlled vocabulary.

[9] https://www.w3.org/RDF/

DRAFT

2 Methodology

This section presents the methodology that we followed in order to develop and publish the GoTriple Vocabulary.

2.1 Choosing LCSH as the base vocabulary

Partners of WP2 agreed that the best approach would be to create a Vocabulary (the GoTriple Vocabulary) based on an existing LOD vocabulary. The first step was to decide which vocabulary to use as the base. To do so, the contributors of T2.4 created a list of existing vocabularies with their main characteristics, such as the number of concepts, supported languages, the domain and whether they cover SSH concepts sufficiently and whether they are published as Linked Open Data. The complete list is shown in Table 1.

Table 1. Cartography of existing vocabularies

Name	URI	License	Features
CESSDA Topic Classification	https://vocabularies.cessda.eu/#!detail/TopicClassification	CC BY 4.0	CESSDA product. - Currently available in: en, fr, de, it and other languages - Main topics only, about 100 concepts - API provided.
European Language Social Science Thesaurus (ELSST)	https://elsst.cessda.eu/id/	bespoke license https://elsst.ukdataservice.ac.uk/elsst-guide/obtaining-elsst.aspx (CC license pending)	-CESSDA product, aimed for international use -Only Social Science (not Humanities) -Currently available in 14 languages: Danish, Dutch, Czech, English, Finnish, French, German, Greek, Lithuanian, Norwegian, Romanian, Slovenian, Spanish, and Swedish -over 3,300 concepts -Available in SKOS format
EuroVoc thesaurus	https://op.europa.eu/en/web/eu-vocabularies/th-dataset/-/resource/dataset/eurovoc	freely available - see https://op.europa.eu/en/web/eu-vocabularies/about	-Owned by European Union -Available in 23 EU languages -Available in SKOS and XML format -over 7,000 concepts -API available
Thesaurus for the Social Sciences (TheSoz)	https://www.gesis.org/en/services/research/tools/thesaurus-for-the-social-sciences	Creative Commons Attribution-Noncommercial-No Derivative Works	- Owned by GESIS - Leibniz Institute for the Social Sciences, a CESSDA member - Available in German, English, French and Russian

DRAFT

		3.0 Germany License.	- 8,000 concepts -Available in SKOS format -aligned to other thesauri but these have not been maintained -API available
UNESCO thesaurus	http://vocabularies.unesco.org/browser/thesaurus/en/	CC-BY-SA 3.0 IGO	-Owned by United Nations Educational, Scientific and Cultural Organization (UNESCO) -available in 4 languages: English, French, Russian and Spanish -4420 concepts -available in SKOS format -API available
ICPSR thesaurus	https://www.vocabularyserver.com/icpsr/	CC BY-NC: Creative Commons Attribution-NonCommercial License	-owned by Inter-University Consortium for Political and Social Research (ICPSR), USA -monolingual (English) -c. 3,500 concepts -API available
Library of Congress Subject Headings	http://id.loc.gov/search/?q= http://id.loc.gov/authorities/subjects.html	public domain data	has to be analysed to extract SSH topics, - hierarchical, - should be quite comprehensive for the humanities, - monolingual (English), yet concrete subject headings mapped to multiple other non-English vocabularies, - multiple mappings (Wikidata, FAST, many national libraries vocabularies; see for example: http://id.loc.gov/authorities/subjects/sh85077507.html) - available in SKOS format, API available
FAST (Faceted Application of Subject Terminology)	http://experimental.worldcat.org/fast/	ODC-By	- available via SKOS,API, - English only - leaner version of LCSH: The purpose of adapting the LCSH with a simplified syntax to create FAST is to retain the very rich vocabulary of LCSH while making the schema easier to understand, control, apply, and use.
Wikidata	https://www.wikidata.org/wiki/Wikidata:Main_Page	CC0	- has to be analysed to extract SSH topics,

DRAFT

			- hierarchical, - multilingual - hub for subject headings identifiers - matched to RDF - rapidly growing in scope and in its use, allow for further connections beyond subject-centred vocabularies (like events etc.)
ERC	https://erc.europa.eu/sites/default/files/document/file/ERC_Panel_structure_2019.pdf		Topic Classification
PACTOLS	http://pactols.frantiq.fr	CC-BY-SA	- Thesaurus for Archaeology - 7 languages : fr, en, es, de, nl, it, ar - 6887 concepts - Owned by FRANTIQ (fr) - SKOS
BNE	http://datos.bne.es/	CC0	- Spanish - SKOS - 372,364 Generic concepts
RAMEAU	https://data.bnf.fr http://rameau.bnf.fr/index_en.html	ETALAB	- french - 224,128 concepts
LEXVO	http://www.lexvo.org/index.html	CC BY-SA 3.0	Languages thesaurus
GEONAMES	https://www.geonames.org/	Commercial	- 25 million geographical names
GEMET (General Multilingual Environmental Thesaurus)	https://www.eionet.europa.eu/gemet/en/themes/		- Multilingual labels - 10442 concepts
COAR Resource Type vocabulary	http://vocabularies.coar-repositories.org/documentation/resource_types/	CC BY 4.0	- Type Classification - Multilingual labels - COAR - mapped with terms and concepts of similar vocabularies and dictionaries
GETTY	http://www.getty.edu/research/tools/vocabularies/	ODC-By 1.0	- Art & Architecture Thesaurus (AAT) - Cultural Objects Name Authority (CONA) - Iconography Authority (IA)™ - Geographic Names (TGN)

DRAFT

			- Categories for the Description of Works of Art (CDWA)
TaDiRAH - Taxonomy of Digital Research Activities in the Humanities	http://tadirah.dariah.eu/vocab/	Open	- 4 languages : en, fr, de, es - Cover : Activities, Objects, Techniques
BBT - Backbone Thesaurus	https://vocabs.dariah.eu/backbone_thesaurus/en/	CC BY 4.0	- 4 languages : en, fr, de, el
Virtual International Authority File	viaf.org	ODC-By 1.0	• 44 active participants/files • 55 million source authority records • 130 million bibliographic records • 256 million links between sources • 30 million external links • 33 million VIAF clusters
OpenEdition Index (v2 in preparation)	https://tools.labocleo.org/indicateurs/indicateurs24.php?shi=49&o=html	(open)	- 175 entries - SSH disciplines, periods, locations - will be publicly available under an open licence in 2020 (in French only) - the vocabulary will be accessible through the OpenTheso software, which provides PIDs, SKOS format and allows for multilingual vocabularies management
BIC Standard Subject Categories	https://ns.editeur.org/bic_categories	open(?)	- 591 entries available - qualifiers for any type of book - superseded by Thema (https://www.editeur.org/151/Thema/)
EKT version of UNESCO Thesaurus (with greek translations, subset)	https://www.semantics.gr/authorities/vocabularies/ekt-unesco/vocabulary-entries/tree?language=en	CC BY-SA 4.0	- The vocabulary comprises the Greek version of the UNESCO Thesaurus terms. It is primarily destined for describing cultural resources, so it focuses on terms within the cultural, educational, social and humanities domain. It is continuously updated and enriched with additional terms on its application. - 1371 terms in english and greek - Created by EKT - LOD - SKOS - API
EKT Thesaurus	https://www.semantics.gr/authorities/vocabularies/g	CC BY-SA 4.0	-7158 terms - Greek labels but for the majority

DRAFT

	eneral-terms-ekt/vocabulary-entries/tree?language=en		there are also English translations - Not only SSH - LOD - SKOS - API
Australian and New Zealand Standard Research Classification (ANZSRC)	https://www.abs.gov.au/ausstats/abs@.nsf/Latestproducts/1297.0Contents12008?opendocument&tabname=Summary&prodno=1297.0&issue=2008&num=&view=	CC BY 4.0	- covers 97 divisions of research - includes divisions in Studies in Human Society (division 13), Language, Communication and Culture (division 20), History and Archaeology (Division 21), Psychology - english language
European Education Thesaurus	https://vocabularyserver.com/tee/en/	open (not defined restrictions)	- 7710 terms - 9 languages (english,french, portugüés, spanish, german, catalan, gaelic, italian, basque) - SKOS - DC - MADS - API

In order to choose a base vocabulary, the following criteria were considered:

- Does the vocabulary cover SSH concepts sufficiently?
- If the vocabulary covers other domains as well, is there a way to extract SSH concepts?
- Is the number of SSH concepts reasonably large? A vocabulary of only a few hundred concepts may not work for annotation. On the other hand, a very large number of concepts would be difficult to manage, especially if they must be enhanced with labels in all our supported languages
- Is the vocabulary published as Open Access, ideally as Linked Open Data?
- Is the vocabulary available in a standard schema, such as SKOS, either via an API or as a downloadable dump?
- Is the vocabulary broadly used and are there existing mappings to national/local vocabularies from which we could obtain labels in our supported languages?

The contributors of task T2.4 agreed on the Library of Congress Subject Heading (LCSH)[10] since it is one of the most widely used vocabularies, it fully covers SSH-related concepts, it is published as LOD in different schemata (SKOS, MADS/RDF) and there are existing mappings to other vocabularies, including Wiki Data[11]. Another advantage of LCSH is that it has already been used by ISIDORE infrastructure and specifically for its annotation service. So it was agreed that the best approach would be to create a Vocabulary based on a subset of LCSH and for partners

[10] https://id.loc.gov/authorities/subjects.html
[11] https://www.wikidata.org/wiki/Wikidata:Main_Page

DRAFT

to add “translations” and/or mappings to national/local vocabularies in order to increase its multilingualism.

LCSH is the list of headings produced from the subject authority file maintained by the United States Library of Congress for use in bibliographic records. It is popularly known by its abbreviation as LCSH. It is a multidisciplinary vocabulary that includes headings in all subjects, from science to religion, to history, social science, education, literature, and philosophy. It also includes headings for geographic features, ethnic groups, historical events, building names, etc. LCSH is the most widely used subject vocabulary in the world. While it provides concepts with labels only in English, parts of it have been translated into numerous languages and/or linked to national/local vocabularies. Today, LCSH is published as LOD and is available in various data models, including SKOS.

2.2 Selecting a reasonable number of SSH concepts from LCSH

LCSH contains more than 242,511 subjects. One challenge was to isolate the SSH-related concepts. Another was to select a proper subset of these. A subset of a few thousand concepts would be large enough for the effectiveness of the annotation service and small enough for enhancing them with labels in the languages supported in GoTRIPLE.

The methodology for selecting the SSH-related concepts was based on identifying basic concepts from the Field of Science and Technology (FOS) classification[12] of the Frascati Manual[13] that cover SSH (14 in total) and then mapping these to 37 broad concepts of LCSH. The mapping is presented in Table 2. Finally, we extracted these LCSH concepts along with their child nodes, using the Linked Data API of the Library of Congress. We ended up with a hierarchical vocabulary of 2565 concepts. The vocabulary is hosted in a platform for managing and publishing LOD Vocabularies developed by EKT, Semantics.gr, under the URI http://semantics.gr/authorities/vocabularies/SSH-LCSH.

TABLE 2. A MAPPING FROM SSH FIELDS OF SCIENCE (FOS) OF FRASCATI TO LCSH

Frascati		LC Subject Headings (LCSH)	
Social Sciences	Psychology	Psychology	http://id.loc.gov/authorities/subjects/sh85108459
	Economics and Business	Economics	http://id.loc.gov/authorities/subjects/sh85040850
		Economic conditions	http://id.loc.gov/authorities/subjects/sh99005736
		Economic history	http://id.loc.gov/authorities/subjects/sh85040817
	Educational sciences	Education	http://id.loc.gov/authorities/subjects/sh85040989
		Music in education	http://id.loc.gov/authorities/subjects/sh85088920
	Sociology	Sociology	http://id.loc.gov/authorities/subjects/sh85124200

[12] https://www.oecd.org/sti/inno/38235147.pdf
[13] Frascati Manual 2015, available at https://www.oecd.org/innovation/frascati-manual-2015-9789264239012-en.htm

DRAFT

	Law	Law	http://id.loc.gov/authorities/subjects/sh85075119
	Political Science	Political science	http://id.loc.gov/authorities/subjects/sh85104440
	Social and economic geography	Economic geography	http://id.loc.gov/authorities/subjects/sh85054016
	Media and communications	Mass media	http://id.loc.gov/authorities/subjects/sh85081863
	Other social sciences	Communication	http://id.loc.gov/authorities/subjects/sh85029027
		Ethnology	http://id.loc.gov/authorities/subjects/sh85045198
Humanities	History and Archaeology	Archaeology and history	http://id.loc.gov/authorities/subjects/sh85006513
		Archaeology	http://id.loc.gov/authorities/subjects/sh85006507
		Classical antiquities	http://id.loc.gov/authorities/subjects/sh85026692
		History	http://id.loc.gov/authorities/subjects/sh85061212
	Languages and Literature	Linguistics	http://id.loc.gov/authorities/subjects/sh85077222
		Classical philology	http://id.loc.gov/authorities/subjects/sh85026710
		Philology	http://id.loc.gov/authorities/subjects/sh85100821
		Philosophy, Ancient	http://id.loc.gov/authorities/subjects/sh85100861
		Literature	http://id.loc.gov/authorities/subjects/sh85077507
	Philosophy, Ethics and Religion	Philosophy	http://id.loc.gov/authorities/subjects/sh85100849
		Religion	http://id.loc.gov/authorities/subjects/sh85112549
		Religious education	http://id.loc.gov/authorities/subjects/sh85112630
		Art and philosophy	http://id.loc.gov/authorities/subjects/sh2002011046
		Folklore	http://id.loc.gov/authorities/subjects/sh85050104
	Arts	Arts	http://id.loc.gov/authorities/subjects/sh85008324
		Art	http://id.loc.gov/authorities/subjects/sh85007461
		Theater	http://id.loc.gov/authorities/subjects/sh85134522
	Other Humanities	Geography	http://id.loc.gov/authorities/subjects/sh85053986
		Library science	http://id.loc.gov/authorities/subjects/sh85076723
		Anthropology	http://id.loc.gov/authorities/subjects/sh85005581
		Civilization	http://id.loc.gov/authorities/subjects/sh85026423

DRAFT

2.3 Using Semantics.gr as the hosting platform

A key requirement for the GoTriple Vocabulary was to be published as a LOD vocabulary in SKOS, which is the standard data model for concept-based vocabularies. This means that the vocabulary itself and all of its concepts should be accessible via persistent URIs both as HTML pages (suitable for humans) and as RDF (suitable for machines). The RDF representation should follow SKOS. Moreover, it should be published under an open license so that anyone can use it

for their own purposes. The latest requirement is very important since the GoTriple Vocabulary is not just an SSH subset of LCSH (which is already important) but it provides labels in 9 additional languages which provides significant added value. Another requirement was that the GoTriple Vocabulary should be continuously updated by the Triple consortium, for example adding/editing labels, adding new links (mappings) to other vocabularies, or adding new concepts.

For all the above reasons, it was necessary that the GoTriple Vocabulary should be hosted and published in a platform dedicated to the development and publication of LOD Vocabularies. It was decided to use Semantics.gr as the hosting platform for the vocabulary.

Semantics.gr [1] is based on a state-of-the-art infrastructure that underpins the development, curation and interlinking of vocabularies, thesauri, classifications and authority files- known collectively as Vocabularies- and their publication as Linked Open Data (LOD). The infrastructure is developed in-house by the National Documentation Centre (EKT). A key feature of Semantics.gr is that, besides SKOS, it can support any Data Model that can be expressed as an OWL ontology or an RDF Schema.

For the initial import of the GoTriple Vocabulary, we developed a batch import Java program that implements the main logic described in section 2.3. The program takes as input a list of LCSH IDs (URIs). It reads each LCSH URI and dereferences it in its RDF/XML representation (SKOS-based application profile). After transforming it in the (SKOS) data model of the GoTriple Vocabulary, it persists it in the GoTriple Vocabulary in semantics.gr. It then reads all its narrower concept URIs (provided in the skos:narrower property). For each “child” URI, it dereferences it in its RDF/XML representation and after transforming it in the data model of the GoTriple Vocabulary, it persists it in the GoTriple Vocabulary under its “parent” concept in semantics.gr. Every GoTriple Vocabulary concept has a URI with the corresponding LCSH id as a suffix and is linked to the corresponding LCSH concept via the skos:exactMatch property (URI Ref).

After the initial import, the GoTriple Vocabulary contained 2565 concepts. Figure 1 illustrates the landing page of the first version of the GoTriple Vocabulary in semantics.gr.

DRAFT

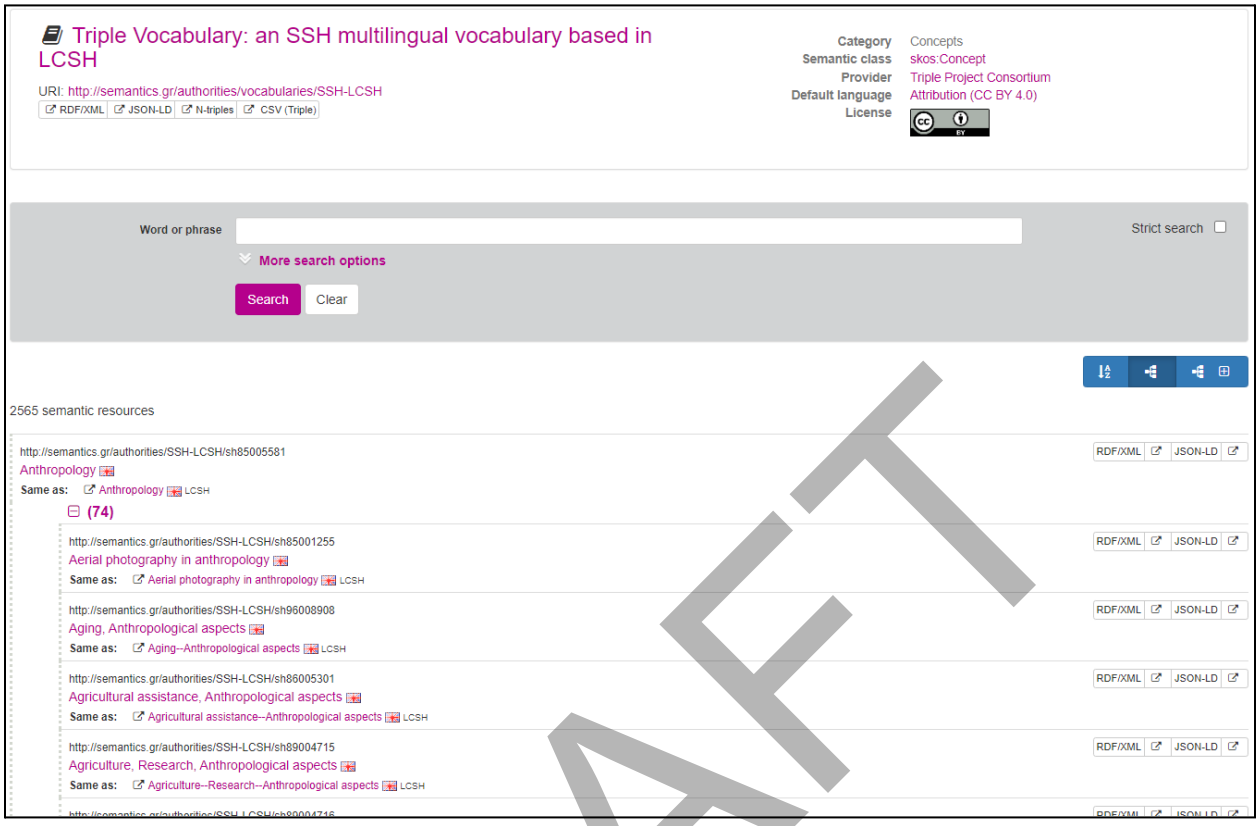


FIGURE 1. The landing page of the 1st version of the GoTriple Vocabulary in semantics.gr

2.4 Making the GoTriple Vocabulary multilingual by leveraging existing mappings to LCSH

Our next action was to make the GoTriple Vocabulary multilingual with respect to our supported languages, since, initially, it had only English labels, like the original LCSH. Our first attempt towards achieving this was to leverage existing mappings between LCSH and other vocabularies.

2.4.1 Adding labels from skos:closeMatch links of LCSH to other vocabularies

First, existing links of LCSH to other vocabularies were imported. Labels in our target languages were then extracted and added in the Vocabulary in the skos:prefLabel property. We developed a java program that implemented the following algorithm:

- It reads each concept of the GoTriple Vocabulary and dereferences its LCSH URI in its RDF/XML representation. From that representation, the skos:closeMatch URI links are extracted.
    - ↺ Each URI link is dereferenced in its RDF/XML representation. From that representation, the skos:prefLabel labels in our supported languages are

DRAFT

extracted. The concept in the GoTriple Vocabulary is enriched with extra skos:prefLabel labels in these languages (unless there are already preferable labels in these languages) and with the corresponding link in the skos:closeMatch property.

2.4.2 Adding labels from the National Library of Florence Vocabulary

Colleagues from ILC created a SPARQL query that returns mappings from URIs of concepts of the National Library of Florence thesaurus to LCSH URIs. First we transformed the SPARQL result to CSV[14]. Then we developed a java program that implemented the following algorithm:

- It reads each LCSH URI in the mapping and if it exists in the GoTriple Vocabulary, the mapped Florence National Central Library URI is dereferenced in its RDF/XML representation.
    - From that representation, the label in Italian is extracted. The concept in the GoTriple Vocabulary is enriched with an extra skos:prefLabel in Italian and with the corresponding link in skos:closeMatch as long as the mapping was not already there (the original LCSH already had some mappings to the National Library of Florence vocabulary).

2.4.3 Adding labels from the Rameau Thesaurus

Colleagues from OpenEdition created a sparql query that returns mappings from URIs of concepts of the Rameau thesaurus[15] to LCSH uris. First we transformed the sparql result to csv[16]. Then we developed a java program that implemented the following algorithm:

- It reads each LCSH URI in the mapping and if it exists in the GoTriple Vocabulary, the mapped Rameau URI is dereferenced in its RDF/XML representation.
    - From that representation, the label in French is extracted. The concept in the GoTriple Vocabulary is enriched with an extra skos:prefLabel in French and with the corresponding link in skos:closeMatch as long as the mapping was not already there (the original LCSH already had mappings to the National Library of Florence vocabulary).

DRAFT

[14] https://docs.google.com/spreadsheets/d/16ISOxNMYqj_oJeuL30kb8CevswmVTiMa3B7TRHGi1TU/edit?usp=sharing
[15] RAMEAU thesaurus: the French subject headings maintained by the National Library of France, available at https://data.bnf.fr/
[16] https://docs.google.com/spreadsheets/d/1r-0_YInvbr90N1PiGDdGQXzOXnz7J8-bh99vRw94qqY/edit?usp=sharing

2.4.4 Adding labels from the ELSST thesaurus

The European Language Social Science Thesaurus (ELSST)[17] is a broad-based, multilingual thesaurus for the social sciences. It is owned and published by the Consortium of European Social Science Data Archives (CESSDA) and its national Service Providers. ELSST is currently available in 16 languages, including English, French, German, Greek and Spanish.

Colleagues from CESSDA created a mapping[18] from selected concepts of ELSST to LCSH. We developed a java program that the following algorithm:

- It reads each LCSH URI in the mapping and if it exists in the GoTriple Vocabulary, the mapped ELSST URI is dereferenced in its RDF/XML representation.
    - From that representation, the labels in French, German, Greek and Spanish are extracted. The concept in the GoTriple Vocabulary is enriched with extra skos:prefLabel in these languages (unless there are already preflabels in these languages) and with the corresponding link in skos:closeMatch.

2.4.5 Validating preferable labels obtained by "skos:closeMatch" mappings

The mappings discussed so far were expressed using the skos:closeMatch property. In the GoTriple Vocabulary, the URI was added in the skos:closeMatch property and the label in the skos:prefLabel, as long as there was not already a skos:prefLabel for that language. However, it was noticed that some of these labels should be removed from the skos:prefLabel property since they were slightly different concepts. In order to correct this, we asked partners to check the skos:prefLabel labels and remove those that were not an exact match. In order to make the process as easy as possible, we created a google spreadsheet[19] with one sheet per language where partners easily validated skos:prefLabel values by simply ticking checkboxes. After this action was completed, labels that were not ticked were removed from the skos:prefLabel property in the GoTriple Vocabulary.

DRAFT

[17] https://www.cessda.eu/Tools/ELSST-Thesaurus

[18] https://docs.google.com/spreadsheets/d/1SHm9JNFsxzvYSHECZ9foMHj9kdep5VerLkJe0jcWwhg/edit?usp=sharing

[19] https://docs.google.com/spreadsheets/d/1nMPWB-3sGxVExozOjTAOV76koPlaKXxWAvjFKeEka2Y/edit?usp=sharing

			Close March (1)			Close March (2)		
URI	Hierarchy	en label	label	URI	pref label ?	label	URI	pref label ?
http://semantics.gr/authorities/SSH-LCSH/sh85000160	▶ Civilization ▶ Learning and scholarship ▶ Humanities ▶ Philosophy ▶ Psychology ▶ Educational psychology	Ability grouping in education	Groupes de niveau	http://data.bnf.fr/ark:/12148/cb11978107x	☑			☐
http://semantics.gr/authorities/SSH-LCSH/sh85000261	▶ Civilization ▶ Learning and scholarship ▶ Humanities ▶ Philosophy	Absurd (Philosophy)	Absurde (philosophie)	http://data.bnf.fr/ark:/12148/cb11947911h	☑			☐
http://semantics.gr/authorities/SSH-LCSH/sh85000353	▶ Linguistics	Acceptability (Linguistics)	Acceptabilité (linguistique)	http://data.bnf.fr/ark:/12148/cb119556629	☑			☐
http://semantics.gr/authorities/SSH-LCSH/sh85000373	▶ Civilization ▶ Learning and scholarship ▶ Humanities ▶ Philosophy	Accidents (Philosophy)	Accident (philosophie)	http://data.bnf.fr/ark:/12148/cb12216126d	☑			☐
http://semantics.gr/authorities/SSH-LCSH/sh85000442	▶ Anthropology ▶ Ethnology	Acculturation	acculturation	http://www.wikidata.org/entity/Q192989	☐	Acculturation	http://data.bnf.fr/ark:/12148/cb13318491g	☑
http://semantics.gr/authorities/SSH-LCSH/sh85000677	▶ Civilization ▶ Learning and scholarship ▶ Humanities ▶ Philosophy	Act (Philosophy)	Acte (métaphysique)	http://data.bnf.fr/ark:/12148/cb12469719w	☑	Philosophie de l'action	http://data.bnf.fr/ark:/12148/cb11937522v	☐
http://semantics.gr/authorities/SSH-LCSH/sh85000892	▶ Civilization ▶ Learning and scholarship ▶ Humanities ▶ Philosophy ▶ Psychology ▶ Personality	Adjustment (Psychology)	Ajustement (psychologie)	http://data.bnf.fr/ark:/12148/cb123451050	☑			☐
http://semantics.gr/authorities/SSH-LCSH/sh85000969	▶ Civilization ▶ Learning and scholarship ▶ Humanities ▶ Philosophy ▶ Psychology	Adolescent psychology	Adolescents -- Psychologie	http://data.bnf.fr/ark:/12148/cb11930829d	☑	Psychologie de l'adolescent	http://www.wikidata.org/entity/Q11943008	☐

FIGURE 2. A google sheet for French where the a colleague from openedition validated prefLabels obtained by "skos:closeMatch" mappings

2.4.6 Language coverage after processing all available mappings

Once the above work to increase the multilingualism of the Triple Vocabulary was completed, we achieved a coverage per language that varied from 5% (Croatian) up to 60% (French), with an average of 20%. Table 3 summarises coverages for each language.

TABLE 3. THE COVERAGE PER LANGUAGE AFTER IMPORTING LABELS OBTAINED BY EXISTING MAPPINGS

Language		Number of concepts with mappings/labels in language (out of 2565)	percentage of concepts with mappings/labels in language (out of 2565)
Greek	el	248	9.67%
French	fr	1624	63.31%
Dutch	nl	339	13.22%
Polish	pl	336	13.10%
German	de	1028	40.08%
Italian	it	686	26.74%
Portuguese	pt	347	13.53%
Spanish	es	406	15.83%
English	en	2565	100.00%
Croatian	hr	153	5.96%
Ukrainian	uk	0	0.00%

DRAFT

2.5 Increasing the multilingualism of the GoTriple Vocabulary using automatic translation

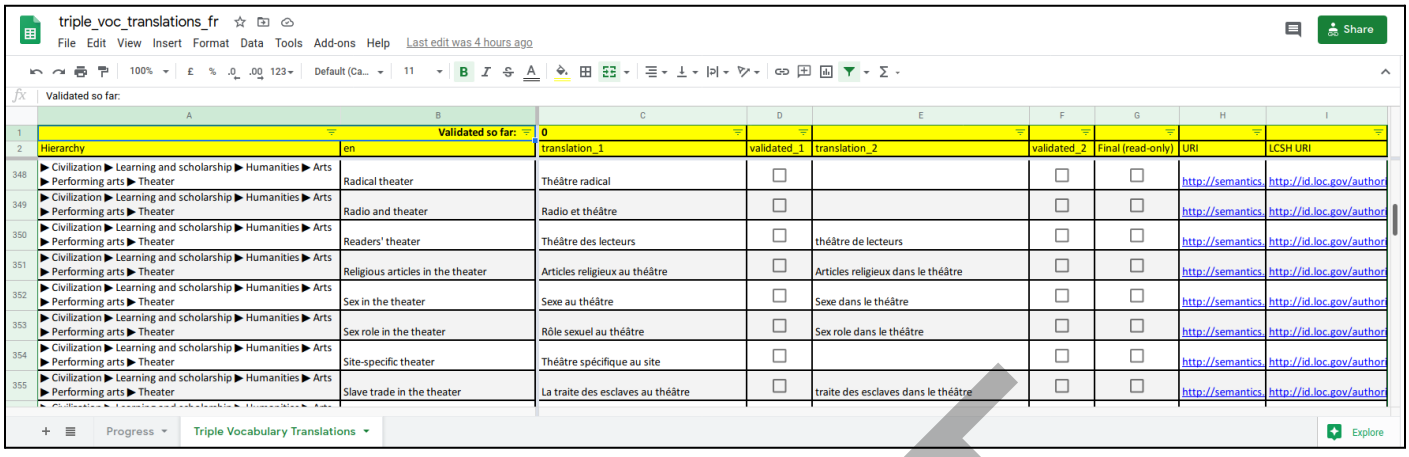
The next challenge was to further increase the language coverage. The initial plan was for partners to add missing labels by manually translating them from English directly in the semantics.gr platform using the edit form (shown in Figure 3). A total of 17 semantics.gr accounts were created in Semantics.gr for members of the project, who were given write access for the GoTriple Vocabulary.

FIGURE 3. Adding more labels for a particular concept using the edit form of Semantics.gr

However, we concluded that manual translation is a labour-intensive process. To overcome this problem, it was suggested that we could use an automatic translation service to generate the missing labels from the English labels. IBL PAN developed a script that, using the Google Translation service, generated all missing labels. Sometimes the google translation service produced 2 candidate translations. Based on these results, we created 9 spreadsheets, one per target language, that included translations in that language[20]. Partners worked on these spreadsheets (each language was assigned to a different partner) in order to validate - and occasionally edit - the automatically generated translations. Figure 4 shows a sample of the G-sheet for French[21] . For each concept, the colleagues responsible for that language inspected "Translation 1" and "Translation 2". They chose the most accurate one by ticking the corresponding checkbox. If needed, they edited or completely overwrote a translation.

DRAFT

[20] https://drive.google.com/drive/folders/1vumdbq3OqfyhV-x0lqJ-Wy0CnG6Buh79?usp=sharing
[21] https://docs.google.com/spreadsheets/d/1qopKZySpH6du9ss0K7wHucII4YzRt3-h4ksPQSeLwSg/edit?usp=sharing



Hierarchy	en	translation_1	validated_1	translation_2	validated_2	Final (read-only)	URI	LCSH URI
▶ Civilization ▶ Learning and scholarship ▶ Humanities ▶ Arts ▶ Performing arts ▶ Theater	Radical theater	Théâtre radical	☐		☐	☐	http://semantics.	http://id.loc.gov/authori
▶ Civilization ▶ Learning and scholarship ▶ Humanities ▶ Arts ▶ Performing arts ▶ Theater	Radio and theater	Radio et théâtre	☐		☐	☐	http://semantics.	http://id.loc.gov/authori
▶ Civilization ▶ Learning and scholarship ▶ Humanities ▶ Arts ▶ Performing arts ▶ Theater	Readers' theater	Théâtre des lecteurs	☐	théâtre de lecteurs	☐	☐	http://semantics.	http://id.loc.gov/authori
▶ Civilization ▶ Learning and scholarship ▶ Humanities ▶ Arts ▶ Performing arts ▶ Theater	Religious articles in the theater	Articles religieux au théâtre	☐	Articles religieux dans le théâtre	☐	☐	http://semantics.	http://id.loc.gov/authori
▶ Civilization ▶ Learning and scholarship ▶ Humanities ▶ Arts ▶ Performing arts ▶ Theater	Sex in the theater	Sexe au théâtre	☐	Sexe dans le théâtre	☐	☐	http://semantics.	http://id.loc.gov/authori
▶ Civilization ▶ Learning and scholarship ▶ Humanities ▶ Arts ▶ Performing arts ▶ Theater	Sex role in the theater	Rôle sexuel au théâtre	☐	Sex role dans le théâtre	☐	☐	http://semantics.	http://id.loc.gov/authori
▶ Civilization ▶ Learning and scholarship ▶ Humanities ▶ Arts ▶ Performing arts ▶ Theater	Site-specific theater	Théâtre spécifique au site	☐		☐	☐	http://semantics.	http://id.loc.gov/authori
▶ Civilization ▶ Learning and scholarship ▶ Humanities ▶ Arts ▶ Performing arts ▶ Theater	Slave trade in the theater	La traite des esclaves au théâtre	☐	traite des esclaves dans le théâtre	☐	☐	http://semantics.	http://id.loc.gov/authori

translation_1	validated_1	translation_2	validated_2	Final (read-only)
Théâtre radical	☐		☐	☐
Radio et théâtre	☐		☐	☐
Théâtre des lecteurs	☐	théâtre de lecteurs	☐	☐
Articles religieux au théâtre	☑	Articles religieux dans le théâtre	☑	☑
Sexe au théâtre	☐	Sexe dans le théâtre	☐	☐
Rôle sexuel au théâtre	☐	Sex role dans le théâtre	☑	☑
Théâtre spécifique au site	☑		☐	☑
La traite des esclaves au théâtre	☐	traite des esclaves dans le théâtre	☐	☐

FIGURE 4. A GOOGLE SPREADSHEET USED FOR VALIDATING FRENCH TRANSLATIONS: VALIDATING A TRANSLATION BY TICKING THE RESPECTIVE CHECKBOX

After the validation process was complete, all validated labels from all spreadsheets were added as skos:prefLabel labels to the GoTriple Vocabulary. This was done automatically by a java application that was developed for that purpose. After the labels were added, the coverage for each language was drastically increased, ranging from 86% (Croatian) to 100% (Ukrainian), as summarised in Table 4.

TABLE 4. COVERAGE PER LANGUAGE AFTER IMPORTING LABELS FROM AUTOMATIC TRANSLATION

Spreadsheet link	Concepts with label in that language obtained by existing mappings	Concepts with automatically generated translation for that language	Validated concepts	New language coverage
Greek	241	2324	2035	**88.73%**
French	1453	1112	766	**86.51%**

DRAFT

Polish	328	2237	2233	**99.84%**
German	888	1677	1673	**99.84%**
Italian	641	1924	1919	**99.81%**
Portuguese	339	2226	2221	**99.81%**
Spanish	397	2168	2164	**99.84%**
Croatian	200	2365	2028	**86.86%**
Ukrainian	0	2565	2565	**100%**

2.6 Adding new concepts from LCSH

After the initial import of the 2565 concepts, we were aware that there were concepts missing.

Colleagues from IBL PAN suggested that 1277 literature terms be added to the GoTriple Vocabulary[22]. Of these, a total of 79 were already in the Triple Vocabulary, reducing the number of new candidate concepts to1198. These concepts fall into one of the following categories:

- **Many of them are narrower concepts of concepts that already exist in the GoTriple Vocabulary.**
  This is a straightforward case. These concepts could be added under existing Triple concepts along with the concepts between them across the hierarchy (intervening concepts). In any case, we had to create an "**inverse import function**" that takes as input a list of LCSH IDs and for each ID (if it does not already exist), it reads its ancestors up to an ancestor that exists in our GoTriple Vocabulary. If such an ancestor exists, the concept and all intervening ancestors are imported.
- **Many of them are "indirectly" narrower concepts of concepts that already exist in the GoTriple Vocabulary.**
  Many of them are not directly narrower concepts of concepts that already exist in the GoTriple Vocabulary but, they (or some of their ancestors) are the "composition" of two or more LCSH concepts, one of them being directly (or indirectly) narrower of concepts that already exist if the GoTriple Vocabulary. This "**composition**" relationship cannot be expressed in SKOS, but it is shown in the concept HTML page in LCSH web ste in the "Components" field. See for example: "Arab countries--Literatures" (https://id.loc.gov/authorities/subjects/sh92005350.html) composed of "Arab countries" (https://id.loc.gov/authorities/subjects/sh85006278.html) and "Literatures" (https://id.loc.gov/authorities/subjects/sh2003003668.html)
  So we decided to extend the "**inverse import function**" to take into account the composition relationship (not only the explicit broader-narrower skos relationship). To

[22] https://docs.google.com/spreadsheets/d/1xq21nKX61WAKis1iuv2ZAfCJZqgCzv0Ng4bGe8ZWoZc/edit#gid=0

DRAFT

do that, and since this 'composition' relationship is not encoded in SKOS, we had to use scraping in the HTML concept pages (using the JSoup library of Java)

- **Many of them do not have a broader relationship (they are so-called orphants) or they are not directly (broader-narrower) or indirectly (via broader-narrower and composition) narrower concepts of concepts that already exist in the GoTriple Vocabulary**.

  We decided not to add these concepts since there is no guarantee that they fall into some broad SSH concept. Also, all these orphan concepts would be added as top-level terms of our GoTriple Vocabulary which was not desirable. Attempting to organise them into hierarchies of our own (which may require adding concepts with no equivalence in LCSH) was not an option, since we did not want to violate the definition of the Triple Vocabulary which " is a subset of LCSH organised in the original hierarchy"

After invoking the "inverse import function" for the suggested LCSH concept URIs, GoTriple Vocabulary was enhanced with 810 new concepts.

Each of these 810 concepts were subject to our standard "label enhancement" data flow, which adds labels in our supported languages leveraging the "skos:closeMatch" links of LCSH to other vocabularies and existing mappings to LCSH. However, all these labels obtained by "skos:closeMatch" links had to first be validated before they were added in the skos:prefLabel (see section 2.4.5). For the remaining missing labels, again we used the google translation service. In all cases, both labels obtained by skos:closeMatch and google translations, had to be validated (and edited if needed), before being added as preferable labels to the Triple Vocabulary. Therefore, we created 9 G-spreadsheets[23], one per target language.

Each g-sheet (for example for language "xx") contains:

- The "hierarchy". The Hierarchy is very important and useful because it easily reveals the context, without having to click on the semantics or LCSH URIs.
- The original label in English (column "en label").
- The automatically generated translations are produced by google translate, in a column named "google translation".
- 2 columns with labels in "xx" produced by "skos:closeMatch" links to "xx" vocabularies or wikidata, named "xx close match (1)" and "xx close match (2)", respectively. Of course these may be empty.
- The URIs of the concept in the GoTriple Vocabulary and in LCSH (columns "Triple Vocabulary URI" and "LCSH URI").

[23] https://drive.google.com/drive/folders/1qG6DtHucAld_w_MiduZCgymXfS9VTiZk?usp=sharing

DRAFT

For each concept, partners were asked to inspect at most 3 labels, in columns “google translation”, “xx close match (1)” and “xx close match (2)” and validate the translation by ticking the appropriate checkbox possibly after editing /correcting the translation.

After the validation process was complete, all validated labels from all G-sheets were added as skos:prefLabel labels to the Triple Vocabulary. This was done automatically by a Java application that was developed for that purpose. After the labels were added, the coverage for each language was drastically increased, ranging from 83% (Ukrainian) to 98% (Spanish), as summarised in Table 5.

Table 5. The coverage per language after importing 810 new concepts

Language		percentage of concepts with mappings/labels in language (out of 3375)
Greek	el	86.64%
French	fr	85.69%
Polish	pl	84.86%
German	de	84.50%
Italian	it	97.01%
Portuguese	pt	88.30%
Spanish	es	98.16%
English	en	100.00%
Croatian	hr	85.39%
Ukrainian	uk	83.91%

DRAFT

3 The GoTriple Vocabulary - An Overview

3.1 The SKOS data model

The data model we chose for the GoTriple Vocabulary is SKOS, a standard for data modelling vocabularies and thesauri. SKOS is an RDF Data Model. It defines classes and properties to describe concepts and their relationship. All concepts or the GoTriple Vocabulary are instances of the skos:Concept class. We used six properties from SKOS: skos:prefLabel, skos:altLabel, skos:browder, skos:narrower, skos:closeMatch and skos:exactMatch. Their semantics and types are summarised in Table 6.

For the purposes of the Triple Project, we created a dedicated schema in Semantics.gr based on SKOS to which only the GoTriple Vocabulary conforms. This way, we can easily add new properties in the future, even custom ones, without affecting other SKOS vocabularies in Semantics.gr.

Table 6. The SKOS properties used in the GoTriple Vocabulary

Class	skos:Concept			
Properties	**qualified name**	**description**	**type**	**multivalued**
	skos:prefLabel	preferable labels (one per language)	Literal	TRUE (one per language)
	skos:altLabel	alternative labels	Literal	TRUE
	skos:browder	URI Ref to broader concept(s) from the same vocabulary. The property defines the hierarchy. Its symmetric property is skos:narrower.	URI Ref	TRUE
	skos:narrower	URI Ref to narrower concept(s) from the same vocabulary. The property defines the hierarchy. Its symmetric property is skos:browder.	URI Ref	TRUE
	skos:closeMatch	URI Refs to similar resources (with similar meaning) of external vocabularies	URI Ref	TRUE
	skos:exactMatch	URI Refs to exact	URI Ref	TRUE

DRAFT

		resources (with exact meaning) of external vocabularies		

3.2 Overview of the Vocabulary

The GoTriple Vocabulary is accessible through its unique permanent URI http://semantics.gr/authorities/vocabularies/SSH-LCSH. Figure 5 illustrates the landing page of the vocabulary. At the top of the page, the user can find some information about the vocabulary, e.g., that the creator is the Triple Consortium, it is based on SKOS and it can be used under the CC BY 4.0 license. Directly below, there is a search box that can be used for searching for specific concepts. And below the search box, all top-level concepts are presented. Figure 6 illustrates all top-level concepts.the user sees all the

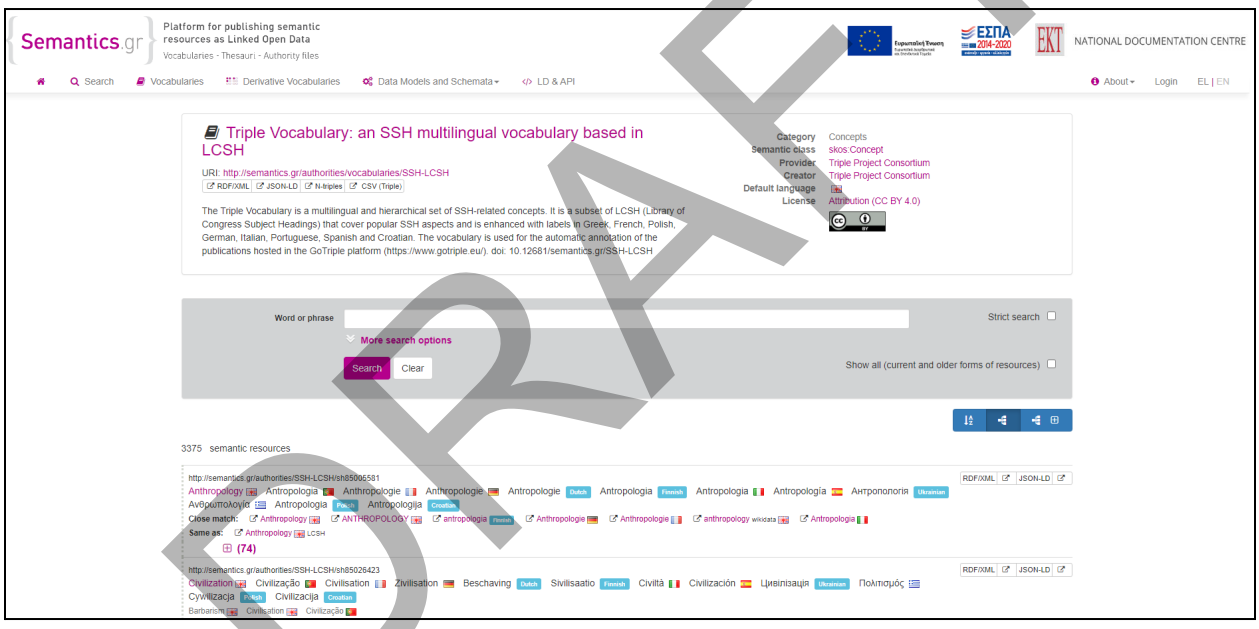


FIGURE 5. The GoTriple Vocabulary landing page in semantics.gr

DRAFT

FIGURE 6. Top-level concepts

DRAFT

The user can click on the '⊞' link to reveal a concept's narrower concepts (Figure 7). To reveal the entire hierarchy they can elect the' ⊞' button. Clicking the' AZ' button allows the user to view the concepts in a list (paginated), as shown in Figure 8.

FIGURE 7. Unfolding concepts to navigate through its narrower concepts.

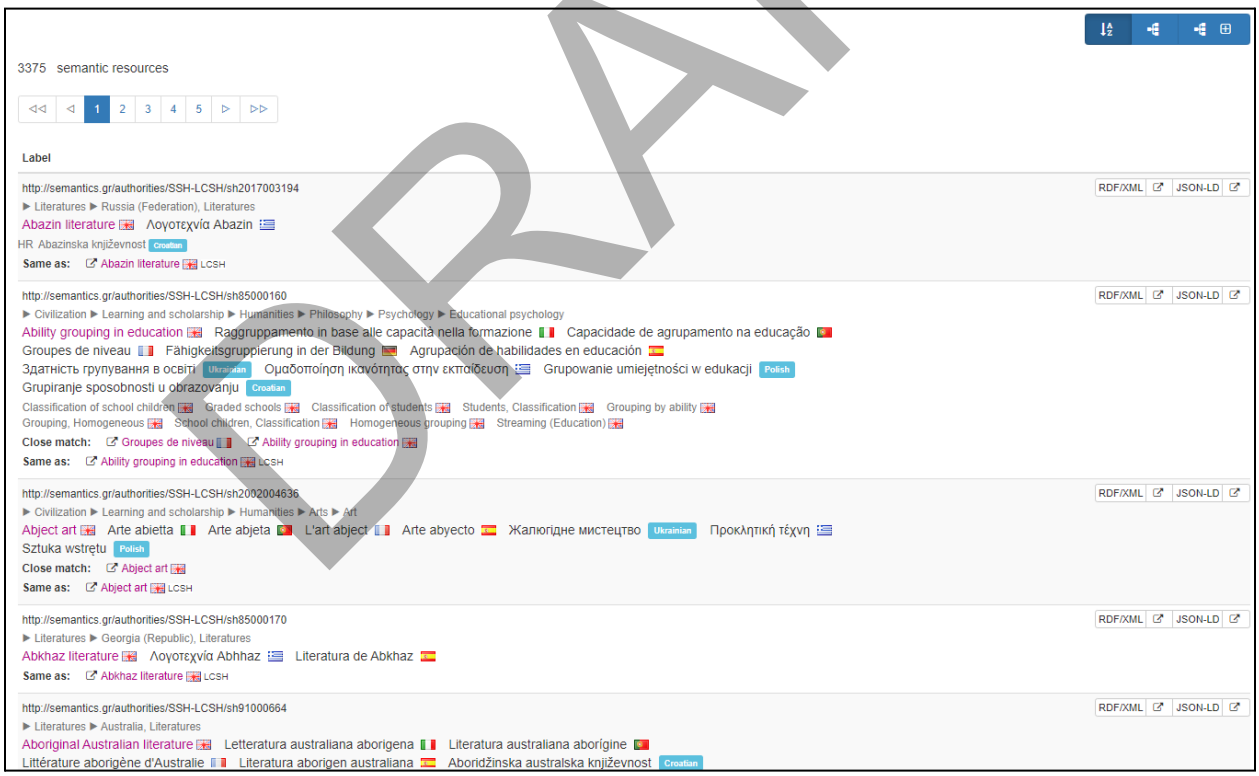


FIGURE 8. The concepts of the vocabulary in list view.

DRAFT

The user can enter a word or phrase in the search box. When they click “search”, only the matching concepts will be returned in a green background (along with all their ancestors if the ‘tree view’ is selected), as shown in Figure 9.

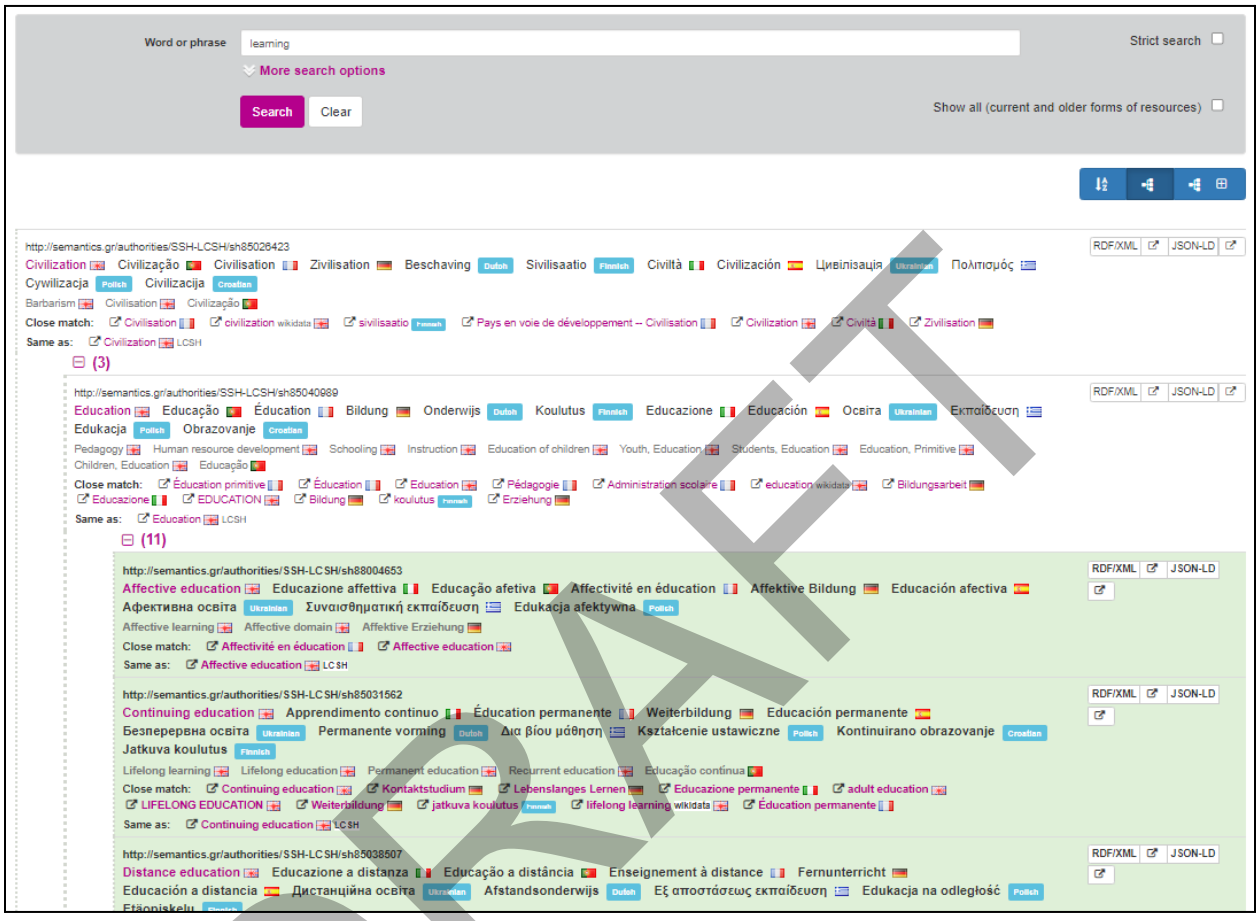


FIGURE 9. Searching concepts by word or phrase.

When the user clicks on a concept, its view page appears that contains all its properties and their values (Figure 10). URI Ref values are shown as links. If a URI Ref value embeds labels in multiple languages as well, these will also appear.

Authorised users that are registered under the ‘Triple Consortium’ provider, can modify a concept using the resource edit form (Figure 11). The edit form is easy to use and includes all the property constraints that have been defined in the vocabulary schema (in case of the GoTriple Vocabulary, the SKOS-based schema).

DRAFT

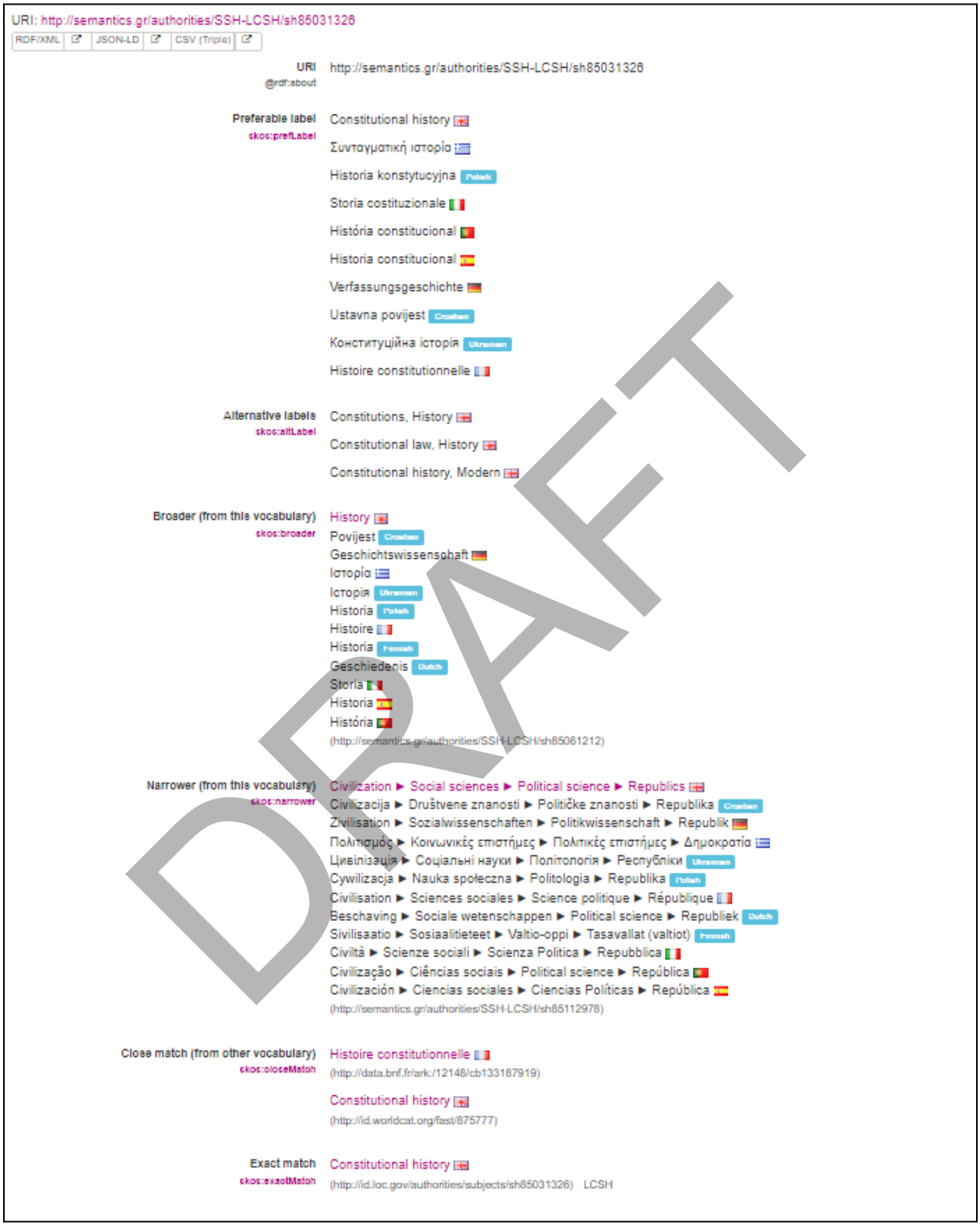


FIGURE 10. A concept view page

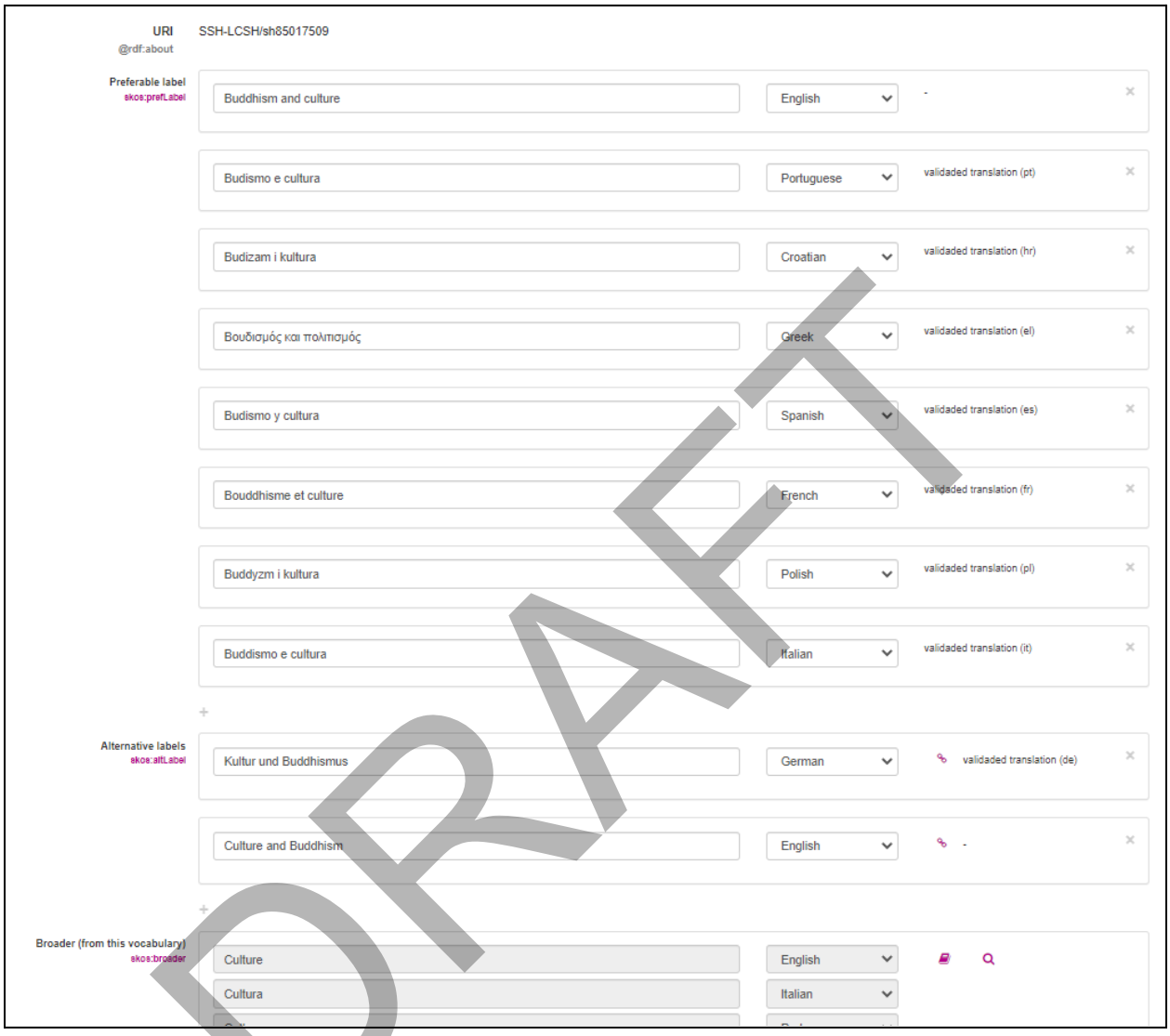


FIGURE 11. The edit form to modify or create a new concept

DRAFT

3.3 Export in RDF and Linked Data endpoint

Semantics.gr publishes both entire vocabularies and their individual semantic resources in RDF using the Data Model defined for each Vocabulary (SKOS in the case of the GoTriple Vocabulary).

Each vocabulary or semantic resourceURI give us its descriptive metadata in 3 different formats using content negotiation:

- When a user enters the vocabulary's URI or the URI of a concept to a common browser they will see the HTML presentation page. The same HTML page will be shown to an

HTTP/REST Client if the command includes the value “text/html” in the header parameter “Accept”.

- If a HTTP/REST Client submits a URI (of the vocabulary or of an individual concept) with the value "application/rdf+xml" or "application/xml" in the header parameter “Accept”, they will get the RDF item description- according to SKOS- in RDF/XML serialization. Alternatively, they can use the ending "/xml" in the URI with no need to define the “Accept” parameter (for example via a common browser).
- Finally, by using the value "application/ld+json" or "application/json" in the Header parameter "Accept", the RDF (modelled in SKOS) description of the item will be viewed in JSON (JSON-LD in particular) serialization. Alternatively, they can use the ending "/json" at the item’s URI with no need to define the "Accept" parameter (for example via a common browser).


```
<rdf:RDF xmlns:ekt="https://www.semantics.gr/authorities/schemanamespaces/ekt#" xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#" xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#" xmlns:skos="http://www.w3.org/2004/02/skos/core#">
 <skos:Concept rdf:about="http://semantics.gr/authorities/SSH-LCSH/sh85017509">
    <skos:prefLabel xml:lang="en">Buddhism and culture</skos:prefLabel>
    <skos:prefLabel xml:lang="pt">Budismo e cultura</skos:prefLabel>
    <skos:prefLabel xml:lang="hr">Budizam i kultura</skos:prefLabel>
    <skos:prefLabel xml:lang="el">Βουδισμός και πολιτισμός</skos:prefLabel>
    <skos:prefLabel xml:lang="es">Budismo y cultura</skos:prefLabel>
    <skos:prefLabel xml:lang="fr">Bouddhisme et culture</skos:prefLabel>
    <skos:prefLabel xml:lang="pl">Buddyzm i kultura</skos:prefLabel>
    <skos:prefLabel xml:lang="it">Buddismo e cultura</skos:prefLabel>
    <skos:altLabel xml:lang="de">Kultur und Buddhismus</skos:altLabel>
    <skos:altLabel xml:lang="en">Culture and Buddhism</skos:altLabel>
    <skos:broader rdf:resource="http://semantics.gr/authorities/SSH-LCSH/sh85034755"/>
   <skos:closeMatch>
     <rdf:Description rdf:about="http://id.worldcat.org/fast/840103">
        <rdfs:label>Buddhism and culture</rdfs:label>
      </rdf:Description>
    </skos:closeMatch>
   <skos:closeMatch>
     <rdf:Description rdf:about="http://data.bnf.fr/ark:/12148/cb12652375q">
        <rdfs:label xml:lang="fr">Bouddhisme et culture</rdfs:label>
      </rdf:Description>
    </skos:closeMatch>
   <skos:exactMatch>
     <rdf:Description rdf:about="http://id.loc.gov/authorities/subjects/sh85017509">
        <rdfs:label xml:lang="en">Buddhism and culture</rdfs:label>
      </rdf:Description>
    </skos:exactMatch>
    <ekt:token>a593cda1-80b7-4a4a-8bdb-d9ec02244309</ekt:token>
  </skos:Concept>
</rdf:RDF>
```


FIGURE 12. The SKOS RDF Representation (in RDF/XML serialisation) of a concept

Apart from content negotiation, a user can access the RDF representation (in RDF/XML or in JSON-LD) by clicking the respective buttons that exist directly below the URI of the vocabulary or of a concept (Figure 12).

DRAFT

4 APPENDIX - MAINTENANCE AND SUSTAINABILITY

The GoTriple Vocabulary may continue to evolve after the Triple Project. Authorised users, perhaps from the OPERAS members, will have edit access in the vocabulary is Semantics.gr. These users can use the functionality of Semantics.gr to insert new concepts or to update existing ones, for example adding new labels or links. However, the further development of the GoTriple Vocabulary should continue to follow the methodology that was followed during the Triple Project. In particular the following requirements should be followed:

- Every new concept should have a direct exact match to an existing concept of LCHS
- Every new concept should be placed in the right branch of the GoTriple vocabulary in alignment with the hierarchy of LCSH. This means that all intervening ancestors should also be inserted.
- For every new concept, an effort should be made to add all Triple languages. When possible, if a label of some language it taken for some external local/national vocabulary, the link should be added in the skos:closeMatch or skos:exactMatch
- All new concepts should belong to the SSH domain

In addition to the above strict requirements, it would be useful if a set of recommendations of good practices were defined, to be taken into account in the future development and maintenance of the GoTriple vocabulary. That was the aim of the document "Recommendations for Implementing Critical Cataloging in the GoTriple Platform"[5] written by Magdalena Wnuk and Svitlana Tarasova. The document is the output of a 5-month process that included the preparation of a literature review and workshop scenario and organisation of two national workshops on critical cataloguing for Polish and Ukrainian stakeholders. Critical cataloging is an interdisciplinary research movement that analyses stereotypes, social and cultural prejudices as well as obsolete or out-of-date information embedded in knowledge organisation systems. Its purpose is not only to identify these biases and distortions, but also to propose solutions aimed at creating the most impartial knowledge organisation systems possible.

The document concluded into the following 8 recommendations:

1. The majority of works and discussions on critical cataloguing take place in North America (USA and Canada). What we need is European perspective on the minority languages, differences in meanings between similar concepts, epochs or historical events in European knowledge organisation systems, Americanisation or other impacts on classifications and subject headings that are subjected to critical cataloguing.

   The GoTriple platform and its vocabulary can provide knowledge on those differences and ambiguities inscribed into European national catalogues interpreting the same historical events, phenomena, cultural heritage and other objects of humanist research. Organising an international event devoted to critical cataloguing of European

DRAFT

vocabularies could be a start for such a debate to develop and engage various disciplinary experts and researchers.

2. There is a need to organise national interdisciplinary as well as expert seminars on selected topics which could bring up issues and help improve inadequate concepts' structure and key words. In some countries people who are professional librarians may not be aware of the terms of critical cataloguing, especially with controlled vocabularies, as their study program may not include it as in the case of Ukraine.

   The GoTriple team has provided a scenario for a critical cataloguing workshop including hands-on activities for the participants (it is an attachment to this document). A workshop can be organised by any party in the respective national community of librarians or researchers interested in reviewing knowledge organisation systems and developing multilingual controlled vocabularies.

3. Engaging the GoTriple users into the process of reviewing and improving the GoTriple controlled vocabulary is not only an inspiring idea but it is also in line with a growing trend in the field of librarianship studies and cataloguing to include users' perspective in the construction of vocabularies.

   The GoTriple team can implement a review mechanism for the vocabulary based on the users' activities on the platform. Users' search phrases and keywords, which are not covered by the GoTriple vocabulary, could be reviewed and added by the decision makers in the Triple Consortium.

4. It would be useful to provide a feedback mechanism through which users could communicate with catalogers suggesting changes to the vocabulary. The Library of Congress Subject Headings (LCSH) as the one responsible for world adopted subject indexing language has a practice of proposals for additions and changes that are looked through by the Policy, Training, and Cooperative Programs Division (PTCP). If they approve them, then an approved list is published (Process for Adding and Revising Library of Congress Subject Headings (loc.gov). The list of already approved proposals is Subject Heading Approved Monthly Lists (classweb.org). Lists are represented monthly since July 15, 2011. As a follow-up to the Polish workshop, a similar form was suggested to be implemented by the National Library of Poland. The IBL PAN team will collaborate on this matter further with representatives of the library's cataloguing team.

   For the GoTriple platform providing users with such a tool seems both feasible and along with the platform's objectives of openness and diversity. It can become a part of the discovery service and be accessible from the research results landing pages. It could be

DRAFT

based on the Google form implemented by the Cataloging Lab on their site: https://cataloginglab.org/suggest-a-heading/.

5. It would be useful if published datasets were linked, for example concepts in the LCSH or RAMEAU link to similar concepts in other catalogues built in other natural languages. That would be helpful for expert users who could easily compare classifications.

   For instance, during the workshop Polish scholars found out that RAMEAU had a larger vocabulary for Romani people than LCSH or Polish National Library catalogue. The same situation was mentioned during the Ukrainian workshop concerning the examples of folklore Cossacks songs and motives. They are usually considered as pheasant ones in LCSH and Ukrainian National Library catalogues whereas in Lviv region there are regional electronic volumes of folk subject headings for this music that apply a wider vocabulary. The catalogues were made on the NGO initiatives of the Lviv Region.

   For starters the GoTriple can provide such case studies as a part of educational material aimed at improving users' skills in methods of critical cataloguing as well as their awareness of knowledge organisation.

6. It is purposeful to analyse ethnic groups and their concepts more attentively and to focus on them whilst creating controlled vocabularies, that will lead to the full cultural layout to be seen in cataloguing. In the case of ethnic studies in Europe, especially those concerning marginalised groups, the insight from both researchers as well as members of those groups is crucial.

   In European countries, such as Ukraine, and in multinational state entities, ethno-national problems have a specificity that extends to social memory which is reproduced in literature. Inclusion of ethnic minorities in the general national historical canon leads to certain complications of categorical status, associated with different influences and significance of these communities for national history and culture. The place of different ethnic groups in the collective memory of various European nations and ethnic groups depend on many factors: population size, place of settlement, ethnic kinship etc.

   During the discussion on Romani subject heading the participants of the workshop taking place in Poland noticed mistakes in the classification of different Romani groups in Europe and the lack of concepts, such as Romaphobia.

DRAFT

During the Ukrainian workshop Ukrainian subject headings describing ethnic minority groups (Boikos, Hutsuls, Lemkos) were discussed. Ethnographers traditionally divide Rusyns into 4 subgroups - Subcarpathian Rusyn, Lemkosv, Boykosv and Hutsuls but those divisions are not easy to draw and therefore cultural heritage of those groups is difficult to classify. Participants pointed out mistakes and misconceptions about the ethnic literature terms and concepts for categorising new types of minority literatures.

The GoTriple team could learn from those examples and critically review these concepts in the GoTriple vocabulary in order to implement the most accurate solutions.

7. Developing advocacy strategies targeted at partners in national communities is another possible direction to take.

   One of the biggest problems at this stage, based on the Ukrainian workshop is that terminological systems of many fields have not yet found their embodiment not only in a systematised presentation within controlled vocabularies, but also in traditional terminological dictionaries. In Ukraine only a part of descriptors is included into the digital version of catalogues while a part of them is still left on paper. The newly added or edited descriptors may play an important role in the search of information, especially its update when it is spoken about such fields as economy, politics, history etc.

   It is to the GoTriple team's vivid interest that such catalogues are digitised fully. Therefore developing advocacy practices and close collaboration with the GoTriple daya providers - libraries and repositories, can be fruitful for the platform's development.

8. Creation of the GoTriple digest could include a critical cataloguing subsection regarding controversial concepts in the GoTriple vocabulary and other systems. This would complement the creation of the user-oriented online form for suggesting changes to the vocabulary (see point 4). Local and national libraries should be added to such a newsletter. Historians, anthropologists, ethnologists, linguists and other interested scholars should be targeted in order to encourage them to take part in vocabulary development.

DRAFT

5 References

[1] Haris Georgiadis, Georgia Angelaki, Elena Lagoudi, Nikos Vasilogamvrakis, Kostas Stamatis, Agathi Papanoti, Alexia Panagopoulou, Katerina Bartzi, Dimitris Charlaftis, Eleni Angelidi, Despina Hardouveli, Poly Karagianni and Evi Sachini, Publishing LOD vocabularies in any schema with Semantics.gr. MTSR 2021 (Preprint available here)

[2] "Resource Description Framework (RDF) Model and Syntax Specification". www.w3.org, available at https://www.w3.org/TR/PR-rdf-syntax/

[3] SKOS Simple Knowledge Organization System Reference, W3C Recommendation 18 August 2009, available at http://www.w3.org/TR/skos-reference

[4] Dewey, Melvil (1876), Classification and Subject Index for Cataloguing and Arranging the Books and Pamphlets of a Library (Project Gutenberg eBook), archived from the original on November 10, 2012, retrieved July 31, 2012

[5] Magdalena Wnuk, & Svetlana Tarasova. (2023). Recommendations for Implementing Critical Cataloging in the GoTriple Platform (Version 1). Zenodo. https://doi.org/10.5281/zenodo.7520453

DRAFT